

Supplementary Materials: Fine-grained Semantic Alignment with Transferred Person-SAM for Text-based Person Retrieval

Anonymous Authors

1 EXPERIMENT DETAILS

1.1 Dataset and Settings

We validate the performance of the proposed method using publicly available datasets that are frequently used in TBPR tasks, including CUHK-PEDES [10], ICFG-PEDES [5], RSPTReid [26] and ATR Dataset [13, 14].

CUHK-PEDES[10] is a classical TBPR dataset with 40,206 images and 80,412 textual descriptions for 13,003 identities. The pedestrian image of CUHK-PEDES comes from five existing person re-identification datasets, CUHK03 [12], Market-1501 [25], SSM [22], VIPER [6], and CUHK01 [11]. The training and testing sets comprise 11,003/1,000 persons with 34,054/3,074 images and 68,108/6,156 sentence descriptions, respectively.

ICFG-PEDES [5] contains 54,522 pedestrian images of 4,102 different identities with more fine-grained text descriptions. The images in ICFG-PEDES are collected from the MSMT17 database [21]. The training set contains 34,674 image-text pairs from 3,102 pedestrians, while the test set contains 19,848 image-text pairs for the remaining 1,000 pedestrians.

RSPTReid[26] is a real scenario text-based person re-identification dataset based on MSMT17 [21]. It contains 20,505 images of 4,101 persons from 15 cameras in total. The training set consists of 3,701 people, 18,505 images, and 37,010 sentence descriptions. The test set includes 1,000 images and 2,000 textual descriptions of 200 pedestrians.

ATR Dataset[13, 14] is a human parsing dataset with 17,706 images and 18 semantic categories. The images in ATR come from diverse sources. Each image is assigned several semantic categories and labeled with fine pixel-level annotations. We randomly split the ATR dataset into the train, valid, and test sets with the ratio of 8 : 1 : 1.

1.2 Implementation Details

We loaded the pre-trained CLIP-B/16 for text and image encoders and randomly initiated the rest of the modules. During training and testing, all images are uniformly scaled to 384×128 , and the maximum length of the text is set to 77. We train the model using the Adam optimizer and set the learning rate of the pre-training module to $1e-5$ with the cosine decay strategy. For the other modules, we set the learning rate $5e-5$. The mask probability p' is set to 0.35. The size of the image patch is 16. For the training of the segmentation model, we use SAM-Base and BERT-base-uncased for the text model. We freeze the base encoder and train only the decoder and other parameters. We set the learning rate to $1e-4$, and train 20 epochs. The image input size of the model is 1024×1024 , and the maximum length of each phrase is 16.

Table 1: The list of semantic merges performed, with the original semantic categories shown on the left and the right side showing which of the other ones we merged these semantic categories with.

Origin	Merged
hair	hair, face
sun-glass	sun-glass, face
left-shoe	left-shoe, right-shoe, left-leg, right-leg

2 DETAILS OF PERSON-SAM TUNING

2.1 Semantic Merging in Person-SAM

We performed a semantic merging on the ATR dataset when training Person-SAM in the main text, and here we explain in detail why we conducted this operation. First, the original ATR data consists of 18 categories (including the background). At the same time, it contains, for example, categories with positional information such as "left-shoe" and "right-shoe," as shown in Fig. 1. These categories usually appear as another entirely different shape due to the masking of the parts. As shown in the left three columns of Figure 1, we demonstrate some masked semantic classes unfavorable for training text-driven semantic segmentation. Therefore, we need to merge these similar semantics to mitigate some of the effects due to masking. In addition to this, for example, as shown in the right two columns of Figure 1, smaller regions like glasses are challenging to achieve detailed linguistic description generation when used alone, whereas the generative model can work well when merged with the face. Finally, as shown in Table 1, we will merge the semantic classes. Instead of merging, we generate the phrases corresponding to these regions separately for the rest of the semantic classes.

In addition to this, inspired by some recent research on visual prompts[4, 19, 23], we still tried to utilize some visual prompts to directly generate descriptions of the corresponding regions, as shown in Fig. 2. We tried to utilize visual prompts for points, regions, and boxes, respectively and asked for the content of the corresponding regions via text. However, this result was not satisfactory, and these attempts invariably produced very noisy results while still generating many errors when confronted with those fine semantic classes. Therefore, we directly filtered as much irrelevant interference as possible during the actual generation process and targeted the prompts for each semantic class.

2.2 Text Prompt for Phrase Generation

After merging the semantic classes and processing the image regions, we next need to design the textual prompts to make the model output fine-grained phrases as correctly as possible. Some existing research also suggests that textual prompts[2, 3, 8, 15, 16, 18], even some punctuation marks that seem small to humans, may contain very unique semantic information to the model. Therefore, we need



Figure 1: When the object is occluded or very small, the generated model can easily misclassify it. The three columns on the left show some examples where BLIP-2 produced an error after the object was occluded. And the two columns on the right show examples where the generated results are erroneous when the objects are very small.

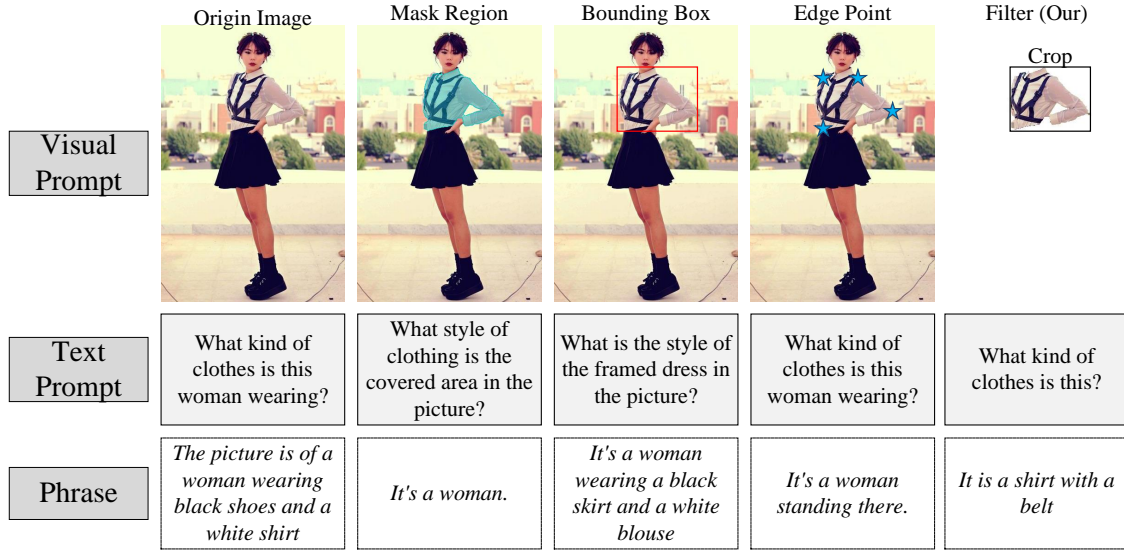


Figure 2: From left to right, visual region processing includes the original image (prompted by text only), mask enhancement, localization box enhancement, keypoint enhancement, and irrelevant region filtering. In several other methods, more or less irrelevant or erroneous bits of content are produced.

to design these linguistic prompts very carefully. We take a natural language perspective, keeping sentences as fluent as possible, and ask questions about the current semantic class. This operation aims to narrow down the model's choices of what to output so that the model outputs more accurate results than without containing the current semantic class word. Taking hair as an example, we show the design of some semantic classes, as shown in Table 2. Finally, we manually evaluate the generation quality of each text prompt on several samples and select the most accurate one for the template. We perform this operation for each semantic class, and finally, we get the specific text prompt templates for all semantic classes as shown in Table 3.

3 ABLATIONS ON IMAGE MASK

Section 4.3 of the main text ablates the attentive cross-modal decoding module. We include an \mathcal{L}_{dec-i} that needs to be mentioned in the main text. The \mathcal{L}_{dec-i} stands for masking the image features and then, in a similar way to the text, improving the model's understanding of the fine-grained features by introducing cross-modal reconstruction. Unlike text, however, current research is still ununified on how to model masked images, so we have tried two main types of approaches, discrete (as used by BeiT [1], for example) and linear (as used by MAE [7], for example).

(1) For discrete, we follow BeiT and quantize the image using a pre-trained VQ-VAE[20] (we chose the DALLE[17] pre-trained

Table 2: Ablation studies on Text Prompt for BLIP-2.

Prompt	Response
<i>None</i>	A black shell.
<i>What object is in this picture? Answer:</i>	This picture has a piece of dark chocolate.
<i>What's in this picture? Answer:</i>	There's a headset in the picture
<i>Question: What's her hair style? Answer:</i>	short
<i>Question: What kind of hair she has? Answer:</i>	She has short, curly hair.
<i>Question: What kind of hair is this? Answer:</i>	The kind of hair she has is short.
<i>Question: What kind of hair is she? Answer: She has</i>	short, curly hair

Table 3: Prompts corresponding to all semantic classes used for BLIP-2.

Semantic	Prompt
hat	Question: What kind of hat is this? Answer: It is
hair	Question: What kind of hair is she? Answer: She has
glasses	Question: What kind of glasses is this? Answer: It is
clothes	Question: What kind of clothes is this? Answer: It is
skirt	Question: What kind of skirt is this? Answer: It is
pants	Question: What kind of pants is this? Answer: It is
dress	Question: What kind of dress is this? Answer: It is
shoes	Question: What shoes is she wearing? Answer: She is wearing
bag	Question: What kind of bag is this? Answer: It is

one), i.e., $\text{ids} = \text{VQ-VAE}(\mathbf{I})$. Then, similarly to text, we mask the image patches to obtain the masked features f'_v . Next, we decode these masked tokens using the textual features f_t , i.e., $\tilde{f}'_v = \text{MCA}(f'_v, f_t, f_t)$. Finally, similarly to text, we find these masked parts and then predict the ids that these features initially corresponded to, i.e., $\mathcal{L}_{dec-i} = \ell_{CE}(\text{MLP}(\tilde{f}'_v), \text{ids})$.

(2) During the experiment, we find that a more significant portion of the vectors acquired by VQ-VAE belong to the background (about 31%, i.e., $\text{ids}_i = 0$). To avoid the category imbalance problem, we ignore these backgrounds and reconstruct only the other tokens $\text{ids}' = \{x \neq 0 | x \in \text{ids}\}$, i.e., $\mathcal{L}_{dec-i} = \ell_{CE}(\text{MLP}(\tilde{f}'_v), \text{ids}')$.

We also have two strategies for linear features: (3) First, the image \mathbf{I} is masked immediately after Patch Embedding and the masked tokens are fed to the subsequent attention layer in ViT.

(4) The other is to mask the image features f_v acquired by ViT. Unlike the discrete strategy, both strategies use MSE loss ℓ_{MSE} to compute the decoding loss \mathcal{L}_{dec-i} . Other than that, the rest of the masking and decoding strategies remain unchanged.

The results of these strategies are shown in Table 4, with strategy (3) achieving the best results. Although image cross-modal decoding alone is effective, performance impairment occurs when combining text cross-modal decoding with image cross-modal decoding, as shown in the main text. How to solve this problem is also part of our subsequent research.

Table 4: Ablations on Image Decoding.

No.	Method	R-1	R-5	R-10
1)	discrete	69.54	85.21	92.65
2)	discrete w/o background	69.77	85.63	92.74
3)	linear & after PE	72.69	86.23	93.67
4)	linear & after ViT	71.56	85.97	93.01

Table 5: Results of Different Person-SAM Text Prompt Methos.

No.	Prompt Feature	mIoU
1)	$f_{A_i}^{[\text{CLS}]}$	56.31
2)	f_{A_i}	59.27
3)	$f_{A_i}^{[\text{CLS}]} \cdot f_t$	58.01
4)	$\theta_{\text{MCA}}(f_p, f_{A_i}, f_{A_i})$	61.55

4 DISCUSSION OF PERSON-SAM STRUCTURE

In the design of Person-SAM, we introduced a text encoder to accommodate text-driven prompts to generate exactly corresponding fine-grained regions. We explored a few ways to use the text features, and the segmentation result on the ATR dataset is shown in Table 5, and we used mAP as the evaluation criterion.

We then briefly describe these methods. **Method** $f_{A_i}^{[\text{CLS}]}$ means using the [CLS] token feature of the text A_i as text prompt. **Method** f_{A_i} means using dense text features (i.e., features of all A_i tokens). **Method** $f_{A_i}^{[\text{CLS}]} \cdot f_t$ means using dot products between dense textual prompts with image features as prompts for the decoder. **Method** $\theta_{\text{MCA}}(f_p, f_{A_i}, f_{A_i})$, which we used, means using Multi-head Cross-Attention(MCA) to obtain the fused textual features between some learnable tokens f_p and the dense textual features as input to the decoder's prompts.

As shown in Table 5, method $\theta_{\text{MCA}}(f_p, f_{A_i}, f_{A_i})$ is better than the other methods, so we use this method in the Person-SAM structure. The reason for this phenomenon is that the fine-grained features require an exact representation, so methods with dense features f_{A_i}

Table 6: Results of Different Alignment Strategy.

No.	Method	R-1	R-5	R-10
	w/o matching	70.42	86.73	92.04
1)	avgPool	32.96(↓)	55.10	65.53
2)	Conv1D	53.87(↓)	76.49	83.35
3)	MHSA	68.99(↓)	87.33	92.07
4)	ELCA (Our)	73.59	89.51	93.55

are better than methods using $f_{A_i}^{[CLS]}$. At the same time, the learnable parameters provide domain adaptation, which helps Person-SAM focus more easily on details related to pedestrians.

5 DISCUSSION ON ALIGNMENT STRATEGY.

In addition to the alignment strategies described in Section ??, inspired by previous work, like GLIP[9], X-VLM[24], we still explored some other alignment strategies. **Method** ‘avgPool’ means using average pooling to aggregate each local feature, i.e., $\bar{f}_v^i = \text{avgPool}(\hat{f}_v^i)$ and $\bar{f}_t^i = \text{avgPool}(\hat{f}_t^i)$, and then employing the InfoNCE loss to constrain the cosine similarity between image and text local features. **Method** ‘Conv1D’ means using 1-d convolution network instead of average pooling in method ‘avgPool’ with the rest remaining unchanged. **Method** ‘MHSA’ means utilizing a multi-head self-attention (MHSA) module to aggregate text and image features, i.e., $\bar{f}_v^i = \text{MHSA}(\bar{v}, \hat{f}_v^i)$ and $\bar{f}_t^i = \text{MHSA}(\bar{t}, \hat{f}_t^i)$, then using the InfoNCE loss to constrain the cosine similarity between the [CLS] feature of \bar{f}_v^i and \bar{f}_t^i , where \bar{v} and \bar{t} are learnable parameter token, $[-]$ is concatenated operation. **Method** ‘ELCA’ is our explicit local concept alignment method described in Section ??.

Table 6 shows the influence of different matching approaches. The benefit of retaining features for all tokens compared to aggregated features, such as ‘avgPool’, suggests that the aggregated features lose some fine-grained information that is trivial in traditional tasks but critical in TBPR. Our ELCA can push the model more strongly to discriminate semantic differences between localizations by interacting with fine-grained features. This result demonstrates the specificity of the TBPR task, i.e., it is a fine-grained task, and the importance of aligning fine-grained semantic features.

6 SCALING ON LARGER MODEL

Table 7: Main result of our SAP-SAM using larger backbone on CUHK-PEDES.

Method	Backbone	R@1	R@5	R@10
Baseline	CLIP (ViT/B-16)	70.42	86.73	92.04
SAP-SAM (Our)	CLIP (ViT/B-16).	75.05	89.93	93.73
Baseline	CLIP (ViT/L-14)	72.13	87.15	92.71
SAP-SAM (Our)	CLIP (ViT/L-14).	76.28	90.87	94.75

We still trained SAP-SAM on a larger backbone network, such as CLIP(ViT/L-14), and the results are shown in Table 7. Our SAP-SAM achieved better results, but we did not use this result in the text for fair comparison.

7 LIMITATIONS

We mainly propose a fine-grained local feature alignment method for images and text to improve the quality of the model’s representation of cross-modal features through fine-grained feature identification and understanding, thus improving the model’s performance in downstream tasks. We focus on some problems in the TBPR task from a fine-grained perspective. However, due to resource constraints, our approach still has some problems:

- In the Person-SAM transfer, due to the limited computational resources, the model we chose is small, which may limit some of the model’s capabilities and lead to less fine-grained results obtained.
- Since there are still some domain differences between the ATR and TBPR datasets in the transfer process, this problem still exists even though we have performed some data style transformations.
- Since we retained all the feature blocks, the computation process took up more time during the learning process of fine-grained features.

In the future, we will also investigate how to learn this relationship faster.

REFERENCES

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [3] Rui Cao, Yihao Wang, Ling Gao, and Meng Yang. 2023. DictPrompt: Comprehensive dictionary-integrated prompt tuning for pre-trained language model. *Knowledge-Based Systems* 273 (2023), 110605.
- [4] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.
- [5] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666* (2021).
- [6] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, Vol. 3. IEEE, 1–7.
- [7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
- [8] Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337* (2022).
- [9] Liunan Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.
- [10] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1970–1979.
- [11] Wei Li, Rui Zhao, and Xiaogang Wang. 2013. Human reidentification with transferred metric learning. In *Computer Vision–ACCV 2012*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 31–44.
- [12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Columbus, OH, USA, 152–159.
- [13] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence* 37, 12 (2015), 2402–2414.

- [14] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2015. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*. 1386–1394.
- [15] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155> 13 (2022).
- [16] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).
- [17] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.
- [18] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.
- [19] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.
- [20] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
- [21] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Salt Lake City, UT, USA, 79–88.
- [22] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2016. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850* 2, 2 (2016), 4.
- [23] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169* (2022).
- [24] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276* (2021).
- [25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. 2015. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171* (2015).
- [26] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 209–217.

523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580